Running head: IMPROVING AAT

Does approaching puppies and avoiding a dead cat improve the effectiveness of approach-avoidance training for changing the evaluation of feared stimuli?

Word count: 5535

Gaëtan Mertens[1], Pieter Van Dessel[2], and Iris M. Engelhard[1]

[1]Department of Clinical Psychology, Utrecht University, Utrecht, the Netherlands

[2]Department of Experiment-Clinical and Health Psychology, Ghent University, Ghent, Belgium

Correspondence concerning this article should be addressed to Gaëtan Mertens, Department of Clinical and Health Psychology, Heidelberglaan 1, room H1.29, Utrecht University, 3584CS Utrecht, the Netherlands.

E-mail: g.mertens@uu.nl

Tel: +31 30 253 75 53

Acknowledgements: The research reported in this paper was funded by a NWO VICI grant (grant number: 453-15-005) awarded to Iris Engelhard and a Postdoctoral fellowship of the Scientific Research Foundation Flanders (FWO-Vlaanderen) awarded to Pieter Van Dessel. We would like to thank Iris Linden, Jodie Molenaar, and Iwan van der Laan for their help collecting the data.

Does approaching puppies and avoiding a dead cat improve the effectiveness of approach-avoidance training for changing the evaluation of feared stimuli?

Word count: 5535

Abstract

Background and Objectives: Approach-avoidance training (AAT) is a procedure for changing people's likes and dislikes that involves executing repeated approach (e.g., pulling a joystick towards yourself) and avoidance actions (e.g., pushing a joystick away from yourself) in response to target stimuli. Typically, this leads to approached stimuli being evaluated more positively than avoided stimuli. However, the evidence that AAT can change evaluations of feared stimuli is mixed. In this preregistered study, we wanted to investigate the effectiveness of a novel version, compared to a more typical version, of AAT for changing the evaluation of fear conditioned stimuli.

Methods: After a differential fear conditioning phase, participants ($N = 80$) were randomly allocated to two conditions: In the novel AAT, participants repeatedly approached one positive picture (i.e., puppies) and avoided one negative picture (i.e., a dead cat) in addition to approaching and avoiding the conditioned stimuli. Participants' evaluations of the stimuli were assessed with explicit ratings and an affective priming task.

Results: We found evidence for the effectiveness of approach-avoidance training to change evaluations of fear conditioned stimuli. However, we found no evidence for the superiority of our novel version of the AAT procedure.

Limitations: The sample size of our study was quite small, limiting the statistical power to detect small effects.

Conclusions: Both a typical and an adjusted version of the AAT procedure proved successful to change conditioned negative evaluations. We compare our findings to previous studies showing limited effectiveness of the AAT procedure with feared stimuli.

Keywords: Approach-avoidance training, evaluations, fear conditioning

Introduction

People's stimulus evaluations are generally predictive for their behavior. For instance, a preference for a certain car manufacturer will be predictive for buying a car from this manufacturer and a preference for social events will be predictive for going to parties and meeting new people. Therefore it is interesting for psychologists, clinicians and marketeers to study stimulus evaluations, including how they are acquired and how they can be changed (De Houwer, Thomas, & Baeyens, 2001; Fazio, Sanbonmatsu, Powell, & Kardes, 1986). One procedure that has shown great promise in its potential to change people's evaluations is approach-avoidance training (AAT). This procedure involves repeatedly performing approach (e.g., pulling a joystick towards oneself) and/or avoid actions (e.g., pushing a joystick away from oneself) in the presence of target stimuli. Prior studies have provided evidence that AAT is effective in changing evaluations for many types of stimuli, including alcoholic beverages, unknown animals, and outgroup faces (Huijding et al., 2009; Phills, Kawakami, Tabi, Nadolny, & Inzlicht, 2011; Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011). However, the effectiveness of AAT has not been demonstrated unequivocally. Other studies have found that AAT does not necessarily add to the effects of other procedures to change evaluations (e.g., Becker, Jostmann, Wiers, & Holland, 2015; Krypotos, Arnaudova, Effting, Kindt, & Beckers, 2015; van Uijen, van den Hout, & Engelhard, 2015).

One important aim of research using AAT and other procedures is to find an efficient means to change pre-existing evaluations. In many cases, particularly in the clinic and for societal relevant topics, it is important that evaluations of stimuli which have a strong prior dislike or preference, such as spiders and sugary beverages, can be modified. Most, if not all,

theories of attitude change predict that changing evaluations of these stimuli will be more challenging than changing evaluations of stimuli which are neutral or ambivalent (De Houwer, 2018; Gawronski & Bodenhausen, 2006; Petty & Cacioppo, 1986).

Changing evaluations appears to be particularly challenging for feared stimuli. That is, in studies investigating fear conditioning, in which initially neutral stimuli (or: conditioned stimuli, CSs) are paired with an aversive unconditioned stimulus (US), procedures which are effective to change conditioned fear responses do not seem to be effective to reduce conditioned evaluative responses (Dirikx, Hermans, Vansteenwegen, Baeyens, & Eelen, 2004; Engelhard, Leer, Lange, & Olatunji, 2014; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010; Luck & Lipp, 2015). The AAT procedure also seems to have limited effect to change conditioned negative evaluations (Krypotos et al., 2015) and to change the evaluation of stimuli with a strong a priori negative valence (e.g., spiders, see van Uijen et al., 2015; though for contrasting evidence see Jones, Vilensky, Vasey, & Fazio, 2013). This difficulty of changing negative evaluations is particularly problematic given that several studies indicate that lingering negative valence may be related to return of fear after successful fear reduction (Dirikx et al., 2004; Kang, Vervliet, Engelhard, van Dis, & Hagenaars, 2018; Zbozinek, Hermans, Prenoveau, Liao, & Craske, 2015). Therefore, ways should be explored to improve the effectiveness of the AAT procedure and other techniques to change evaluations.

AAT may be improved based on predictions of cognitive theories of AAT effects (Krishna & Eder, 2019). According to one recent theory, AAT changes evaluations due to the specific inferences that participants make based on the actions they perform in relation to the stimulus (Van Dessel, Hughes, & De Houwer, 2018). For instance, executing repeated avoidance

actions in the presence of alcohol-related stimuli could lead to the inference that alcohol should be avoided (and this inference may guide future behavior). Importantly, approach and avoidance actions do not always provide clear cues about the specific inferences that should be made, which may explain some of the contrasting findings in the literature. As noted in Van Dessel et al. (2018), one crucial inference that participants need to make in order to show AAT effects might relate to the relation between the performed action and valence. For instance, to infer that a stimulus is positive after learning that one has approached the stimulus, it might be important to realize that approach is a positive action or that people typically approach positive stimuli. From this perspective, disambiguating the evaluative connotation of approach and avoidance actions may help to improve the effectiveness of AAT.

One potential way to achieve this is by manipulating the type of other (non-focal) stimuli that participants have to approach and avoid. In a recent study, we found that AAT effects for target stimuli depended on the identity of non-focal stimuli that participants had to approach and avoid (Mertens, Van Dessel, & De Houwer, 2018). Participants approached stimuli that were previously conditioned with an electric shock and avoided a safety stimulus (or vice versa), while also approaching and avoiding neutral target stimuli (i.e., non-words). When participants approached the safety stimulus and avoided the conditioned stimulus, the approached non-word became more positive and the avoided non-word became more negative (i.e., the typical AAT effect). However, the pattern reversed when participants approached the conditioned stimulus and avoided the safety stimulus, such that the approached non-word now became more negative and the avoided non-word more positive (directly contrasting typical approach-avoidance training effects). Presumably, participants in the latter condition considered the approach action

as negative rather than positive because they repeatedly approached a negative stimulus (i.e., a conditioned stimulus which was previously paired with a shock). Note that the same reasoning might (partially) explain why AAT effects have limited effects for changing evaluations of feared stimuli (i.e., approached actions are considered negative due to repeatedly approaching a feared stimulus).

In the current study, we wanted to exploit the effects of approaching and avoiding non-focal stimuli to improve AAT effects for stimuli with a strong a priori valence. Therefore, we included one highly positive stimulus (i.e., a picture of puppies) which participants had to approach in addition to the target stimulus, and one highly negative stimulus (i.e., a picture of a dead cat) which participants had to avoid in addition to another target stimulus. As target stimuli, we used two neutral female faces of which one was conditioned in a preceding phase using an aversive loud sound. Participants in the control AAT condition had to approach and avoid neutral stimuli (i.e., a picture of a clock and a picture of an umbrella) besides approaching and avoiding the target stimuli. We predicted that including a positive and negative picture in the AAT procedure, which participants had to approach and avoid respectively, would result in larger shifts in valence for the target stimuli compared to the control condition. Additionally, we investigated whether our novel AAT procedure could prevent the re-acquisition of conditioned valence and fear.

Method

Pre-registration

The sample size, design, procedure and data analyses steps were pre-registered on the Open Science Framework prior to the data collection (https://osf.io/wphq8/).

Participants

Eighty students (17 males, 63 females; mean age = 21.65, SD = 2.38) participated in exchange for €4 or course credit. This sample size provided good statistical power (> .99) to detect moderately sized interactions (Cohen's f = .25) between within-subjects and between-subjects factors (Faul, Erdfelder, Lang, & Buchner, 2007). Participants were recruited through flyers and posters on campus and were screened for physical and mental health. All participants completed an informed consent form and were instructed that they could discontinue the experiment at any point without any negative consequences.

Material

Apparatus. The experiment was programmed using Inquisit (https://www.millisecond.com/). Approach and avoidance actions were executed with a Wingman Attack 3 joystick.

Stimuli. The unconditioned stimulus (US) was a 1000 ms white noise sound of approximately 95 dB presented through Sennheiser headphones. Conditioned stimuli (CSs) were two neutral female faces taken from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) presented on a 23 inch screen with a resolution of 1920 by 1080 pixels. One positive, one negative and two neutral pictures were taken from the International Affective Pictures System (IAPS; Lang, Bradley, & Cuthbert, 2008). The positive stimulus was a picture of puppies (IAPS picture 1710; valence: M = 8.59, SD = 0.99) and the negative stimulus was a picture of a dead cat (IAPS picture 9571; valence: M = 1.38, SD = 1.09). The neutral pictures were a picture of a lamp

(IAPS picture 7175; valence: $M = 4.95$, $SD = 0.80$) and an umbrella (IAPS picture 7150; valence: $M = 4.69$, $SD = 1.19$).

Procedure

Acquisition phase. Participants were placed in a soundproofed room behind a computer monitor with the keyboard and joystick attached. The computer task started with the presentation of the two neutral faces. Participants had to rate the valence and their fear of each picture on a slider (0 = *Very negative/Not anxious*, 50 = *Neutral*, 100 = *Very positive/Very anxious*; pre-acquisition ratings). During the Acquisition phase, participants were shown one face at a time. One of the two faces (counterbalanced across participants) was followed by the US (CS+), while the other face was never followed by the US (CS-). The CSs were presented for 5 s with an inter-trial-interval (ITI) of 11, 12, or 13 s for 16 trials. After the Acquisition phase, the faces were rated on valence and fear again (post-acquisition ratings). Furthermore, participants had to indicate whether they had noticed a relationship between the faces and the noise (contingency awareness) and which face was followed by the noise (contingency check).

Approach-avoidance training. Participants in the valenced AAT condition were instructed to approach the picture of the puppies and the CS+ face by pulling the joystick towards them. On the other hand, they had to avoid the picture of the dead cat and the CS- face by pushing the joystick away. Participants in the neutral AAT condition were instructed to do the same, except that they had to approach the picture of the lamp instead of the puppies and had to avoid the picture of the umbrella instead of the dead cat. The AAT consisted of 96 trials with a break after 48 trials. Pictures remained on the screen until participants executed a valid response

(Phills et al., 2011). ITI was set at 11 s. A red cross appeared for 500 ms when participants gave an incorrect response. No zoom effect (e.g., Wiers et al., 2011) or perspective grid (e.g., Jones et al., 2013) were applied. The relation of the joystick action (pulling/pushing) to approach/avoidance was disambiguated to the participants by the provided instruction at the beginning of the task (i.e., pull the joystick towards you and push the joystick away from you). At the end, participants had to rate valence and fear of each CS face (post-AAT ratings).

Affective priming task. Next, participants completed the affective priming task. They were instructed to indicate whether a given word had a positive (e.g., happy, holiday, love) or negative (e.g., war, murder, hate) valence by pressing the "i" or "e" key, respectively, as fast and accurate as possible. One trial consisted of a fixation cross for 500 ms, a black screen for 500 ms, and a CS for 200 ms. Hereafter a positive or negative word was presented. ITI was 500, 1000, or 1500 ms. The task started with a practice block of 10 trials in which a neutral prime ("$ùµ=:#") was given instead of a CS. Incorrect responses were followed by a red cross for 2 s. The main task consisted of 80 trials in which no feedback was given. Acquired CS valence can be distilled from categorization speed, because speed is mediated by the prime's valence.

Re-acquisition. Re-acquisition of fear was tested after two presentations of the CS- face and two presentations of the CS+ face followed by the US (identical to the acquisition phase). Participants then rated valence and fear of the CS faces again (post-reacquisition ratings). Finally, participants filled in the Dutch translation of the Life Events Checklist (Gray, Litz, Hsu, & Lombardo, 2004). Results of this questionnaire were used for exploratory purposes.

Data preprocessing and analysis

Participants who responded incorrectly on the contingency awareness and contingency check questions were excluded from the analyses because contingency awareness is considered a prerequisite for conditioning to take place (e.g., Mertens, Wagensveld, & Engelhard, 2019). All analyses were also performed on the data of all 80 participants. These analyses provided essentially the same results. However, when results with the full dataset differed, this is indicated and explained in a footnote. Several repeated measures ANOVA's were run to test the effectiveness of our interventions on explicit stimulus evaluations with Stimulus (CS+ and CS-) and Time (pre and post-ratings) as within-subjects factors and Condition (neutral AAT and valenced AAT) as a between-subjects factor. First, the success of the acquisition phase was tested by comparing stimulus evaluations pre and post-acquisition. Second, the effectiveness of the AAT was tested by comparing the stimulus evaluations post-acquisition and post-AAT. Third and finally, stimulus evaluations post-AAT and post-reacquisition were compared. The results of the AP task was analyzed with a repeated measures ANOVA with Prime stimulus (CS+ or CS-) and Target type (positive or negative) as within-subjects factors, and Condition (neutral AAT and valenced AAT) as a between-subjects factor. An alpha level of .05 is applied for all analyses.

Results

Final sample

Based on the contingency awareness question, four participants were excluded from the analyses because they indicated that they did not notice the relationship between the faces and the noise. Based on the contingency check question, two additional participants were excluded from the analyses because they failed to identify the correct face which was paired with the

noise. Therefore, the final sample consisted of 74 participants (valenced AAT condition: $n = 36$, neutral AAT condition: $n = 38$).

Valence ratings

Prior to analyzing each phase separately, we conducted an omnibus ANOVA to establish that our manipulations in each phase affected valence ratings and to control for the overall alpha-level. Specifically, a repeated measures ANOVA with within-subjects factors Time (baseline, post-acquisition, post-AAT, and post-reacquisition) and Stimulus (CS+, CS-), and between-subjects factor Condition (neutral AAT and valenced AAT) was conducted. This ANOVA showed a main effect of Time, $F(3, 216) = 9.39$, $p < .001$, $\square^2_p = .12$, and Stimulus, $F(1, 72) = 61.36$, $p < .001$, $\square^2_p = .46$, and a two-way interaction between Time and Stimulus, $F(3, 216) = 40.01$, $p < .001$, $\square^2_p = .36$. The expected three-way interaction between Time, Stimulus, and Condition was not significant, $F(3, 216) = 0.82$, $p = .485$, $\square^2_p = .01$. This ANOVA is further deconstructed in the following paragraphs.

Acquisition. The repeated measures ANOVA with pre and post-acquisition valence ratings showed a significant main effect for Stimulus, $F(1, 72) = 44.21$, $p < .001$, $\square^2_p = .38$, and a significant interaction effect between Stimulus and Time, $F(1, 72) = 62.20$, $p < .001$, $\square^2_p = .46$. Paired sample t-tests showed that the CS- became more positive from before acquisition to after acquisition, $t(73) = 5.73$, $p < .001$, $d_z = 0.67$ (see Figure 1). The CS+ became more negative from before to after the acquisition phase, $t(73) = -7.17$, $p < .001$, $d_z = -0.83$ (see Figure 1). This did not differ between the conditions ($F < 1$).

Approach-avoidance training. The repeated measures ANOVA with post-acquisition and post-AAT valence ratings showed a significant main effect for Stimulus, $F(1, 72) = 52.44$, $p < .001$, $\square^2_p = .42$, and Time, $F(1, 72) = 22.97$, $p < .001$, $\square^2_p = .24$. Also, a significant interaction effect between Stimulus and Time was found, $F(1, 72) = 39.61$, $p < .001$, $\square^2_p = .36$. Paired sample t-tests indicated slightly lower positive valence scores for the CS- after the AAT compared to before, $t(73) = -3.00$, $p = .004$, $d_z = -0.35$. For the CS+, scores increased from before to after the AAT, $t(73) = 7.16$, $p < .001$, $d_z = 0.83$, reflecting a more positive valence towards the CS+ (see Figure 1). There were no main or interaction effects with Condition, Fs(1, 72) < 1.47, ps > .230, $\square^2_p$s < .02. The interaction between Stimulus and Time was found for both groups (neutral AAT: $F(1, 37) = 10.04$, $p = .003$, $\square^2_p = .21$; valenced AAT: $F(1, 35) = 41.56$, $p < .001$, $\square^2_p = .54$). This indicates that AAT shifted valence ratings in both conditions, though the effect is somewhat more pronounced in the Valenced AAT group (see Figure 1; and see the Supplementary Materials for the precise means and standard deviations for CS+ and CS- valence and fear ratings in all phases of the experiment).

Re-acquisition. The repeated measures ANOVA with post-AAT and post-reacquisition valence ratings showed a significant main effect for Stimulus, $F(1, 72) = 46.55$, $p < .001$, $\square^2_p = .39$, and for Time, $F(1, 72) = 21.24$, $p < .001$, $\square^2_p = .23$. Also, a significant interaction effect was found between Stimulus and Time, $F(1, 72) = 51.53$, $p < .001$, $\square^2_p = .42$. Paired sample t-tests showed that for the CS-, valence ratings increased from before to after re-acquisition, $t(73) = 3.88$, $p < .001$, $d_z = 0.45$. For the CS+, valence ratings significantly decreased from before to after re-acquisition, $t(73) = -7.67$, $p < .001$, $d_z = -0.89$, indicating that the CS+ was evaluated as

more negative after re-acquisition (see Figure 1). There were no main or interaction effects with Condition, Fs(1, 72) < 2.1, ps > .160, $\square^2_p$s < .03

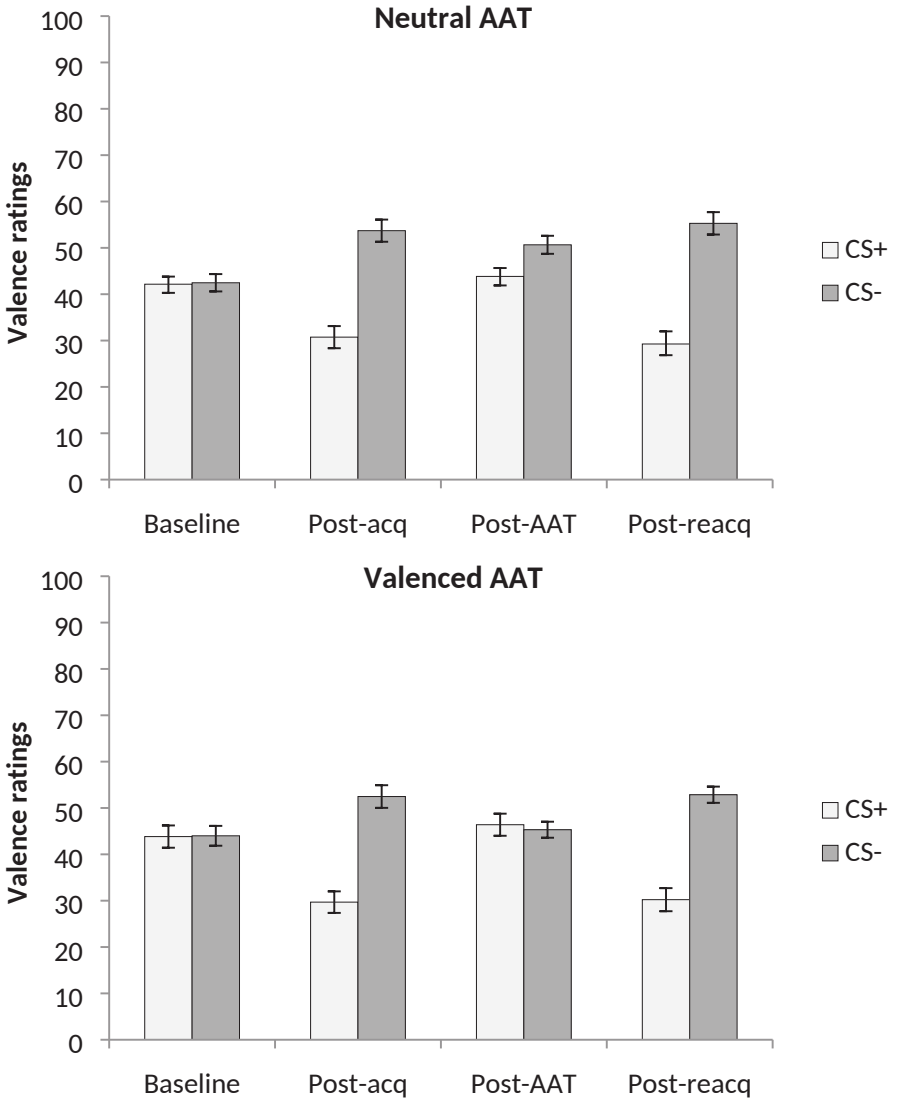

Figure 1. Valence ratings throughout the different phases of the experiment (baseline measurement, post-acquisition, post AAT and post-reacquisition) in the two different conditions of the experiment (AAT + approaching neutral pictures [Neutral AAT] and AAT + approaching positive and negative pictures [Valenced AAT]).

Fear ratings

As for the valence ratings, prior to analyzing each phase separately, we conducted an omnibus ANOVA to establish that our manipulations in each phase affected fear ratings and to control for the overall alpha-level. This ANOVA showed a main effect of Time, $F(3, 216) = 9.95$, $p < .001$, $\eta^2_p = .12$, and Stimulus, $F(1, 72) = 123.64$, $p < .001$, $\eta^2_p = .63$, and a two-way interaction between Time and Stimulus, $F(3, 216) = 51.17$, $p < .001$, $\eta^2_p = .42$. The expected three-way interaction between Time, Stimulus, and Condition was not significant, $F(3, 216) = 2.08$, $p = .103$, $\eta^2_p = .03$. This ANOVA is further deconstructed in the following paragraphs.

Acquisition. The repeated measures ANOVA with pre and post-acquisition fear ratings showed a significant main effect for Stimulus, $F(1, 72) = 100.33$, $p < .001$, $\eta^2_p = .58$ and Time, $F(1, 72) = 19.17$, $p < .001$, $\eta^2_p = .21$. A significant interaction between Stimulus and Time was found, $F(1, 72) = 88.97$, $p < .001$, $\eta^2_p = .55$. A paired sample *t*-test showed that participants found the CS- less fearful after acquisition compared with before, $t(73) = -4.43$, $p < .001$, $d_z = -0.52$ (see Figure 2). Conversely, participants rated the CS+ as more fearful after compared to before acquisition, $t(73) = 8.86$, $p < .001$, $d_z = 1.03$ (see Figure 2). There were no main or interaction effects with Condition, Fs$(1, 72) < 3.70$, ps $> .058$, η^2_ps $< .05$[1].

Approach-avoidance training. The repeated measures ANOVA with post-acquisition and post-AAT fear ratings showed a significant main effect for Stimulus $F(1, 72) = 103.77$, p

[1] The three-way interaction between Stimulus, Time, and Condition, $F(1, 78) = 3.95$, $p = .050$, $\eta^2_p = .048$, as well as the main effect of Condition, $F(1, 78) = 3.99$, $p = .049$, $\eta^2_p = .049$, were significant when all eighty participants were included in the analysis. This suggests a difference in acquisition between the valenced AAT and neutral AAT conditions that can only be a chance effect given that the crucial manipulation that differs in the two conditions was not yet administered in the acquisition phase.

< .001, $\square^2_p$ = .59, and Time $F(1, 72) = 21.53$, $p < .001$, $\square^2_p$ = .23. Also, interaction effects were found between Stimulus and Time, $F(1, 72) = 51.00$, $p < .001$, $\square^2_p$ = .42, and Stimulus and Condition, $F(1, 72) = 5.44$, $p = .022$, $\square^2_p$ = .07. The interaction between Stimulus and Condition was due to higher fear ratings to the CS+ in the neutral AAT condition compared to the valenced AAT condition, while CS- fear ratings were comparable in the valenced AAT condition and neutral AAT condition. Importantly, the interaction between Stimulus and Time was due to a slight increase in fear ratings for the CS- from before to after AAT, $t(73) = 2.54$, $p = .013$, d_z = 0.30 (see Figure 2). On the other hand, CS+ fear ratings significantly decreased from before to after AAT, $t(73) = -8.46$, $p < .001$, $d_z = -0.98$ (see Figure 2). However, this effect did not differ between the conditions since no three-way interaction with Condition was found ($F < 1$). The interaction between Stimulus and Time was found for both groups (neutral AAT: $F(1, 37)$ = 28.37, $p < .001$, $\square^2_p$ = .43; valenced AAT: $F(1, 35) = 23.21$, $p < .001$, $\square^2_p$ = .40).

Re-acquisition. The repeated measures ANOVA with post-AAT and post-reacquisition fear ratings showed a significant main effect of Stimulus, $F(1, 72) = 74.84$, $p < .001$, $\square^2_p$ = .51, and for Time, $F(1, 72) = 15.59$, $p < .001$, $\square^2_p$ = .18. Also a significant interaction between Stimulus and Time, $F(1, 72) = 54.36$, $p < .002$, $\square^2_p$ = .43, and a significant interaction between Stimulus and Condition[2], $F(1, 72) = 4.59$, $p < .036$, $\square^2_p$ = .06, were found. The interaction between Stimulus and Condition was due to higher fear ratings to the CS+ in the neutral AAT condition compared to the valenced AAT condition. CS- fear ratings were comparable in the valenced AAT condition and neutral AAT condition (see Figure 2). Paired sample *t*-tests showed

[2] The Stimulus*Condition interaction was nonsignificant when all eighty participants were included in the analysis, $F(1, 78) = 3.83$, $p = .054$, $\square^2_p$ = .05.

that CS- fear ratings decreased over time, $t(73) = -2.68$, $p = .009$, $d_z = -0.31$ (see Figure 2). On the other hand, fear ratings for the CS+ significantly increased from before to after reacquisition, $t(73) = 7.04$, $p < .001$, $d_z = 0.82$ (see Figure 2). The interaction effect between Time and Condition, and between Stimulus, Time, and Condition were not significant, $Fs(1, 72) < 1$, $ps > .330$, $\square^2_p s < .02$.

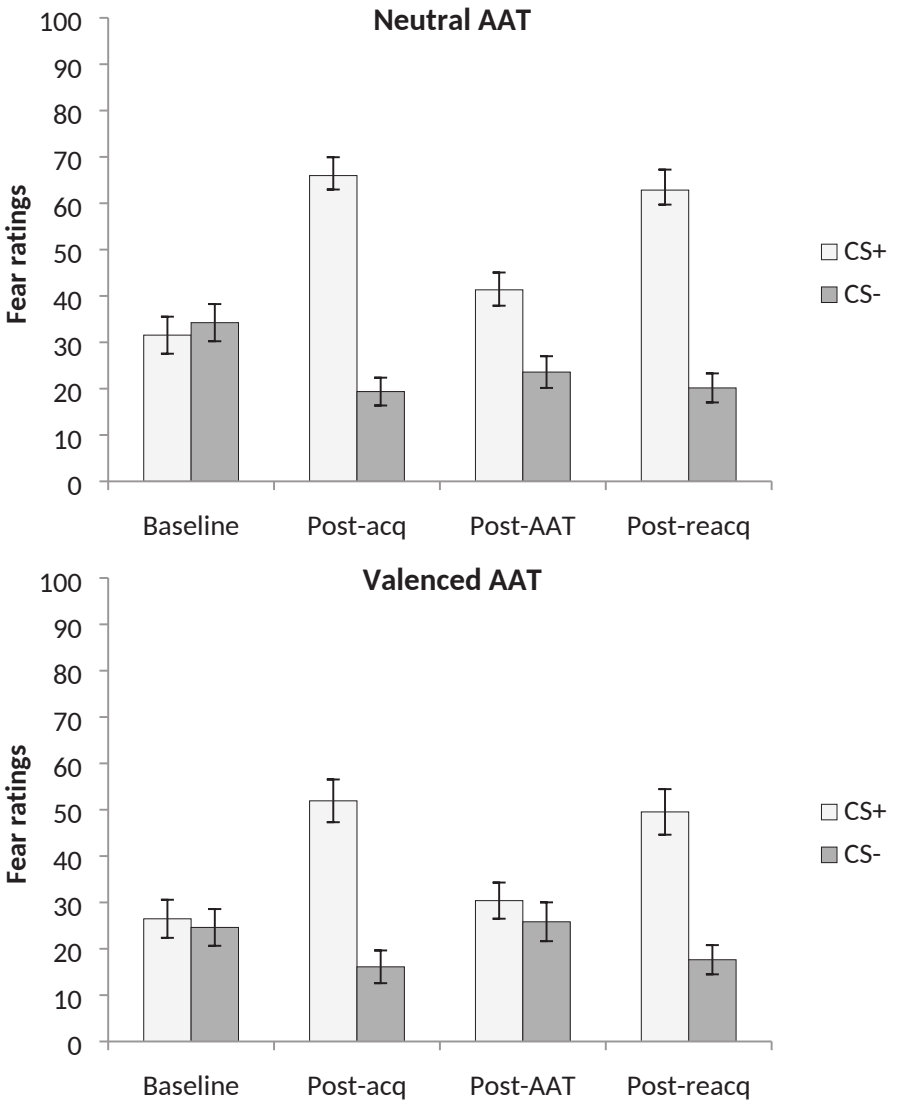

Figure 2. Fear ratings throughout the different phases of the experiment (baseline measurement, post-acquisition, post AAT and post-reacquisition) in the two different conditions of the experiment (AAT + approaching neutral pictures [Neutral AAT] and AAT + approaching positive and negative pictures [Valenced AAT]).

AP task

A repeated measures ANOVA[3] on the reaction times (RTs; only trials with a correct response) of the AP task with Target (positive, negative) and Prime (CS+, CS-) as within-subject factors and Condition (valenced AAT, neutral AAT) was conducted. The only significant effect in this analysis was the two-way interaction between Target and Prime, $F(1, 72) = 5.89$, $p = .018$, $\eta^2_p = .08$. This interaction was due to faster RTs to the CS- when primed with a positive word ($M = 652$, $SD = 135$) than primed with a negative word ($M = 683$, $SD = 165$), $t(73) = 3.18$, $p = .002$, $d_z = 0.37$. The reverse was true for the CS+ (positive prime: $M = 676$, $SD = 199$; negative prime: $M = 665$, $SD = 159$), $t(73) = -0.89$, $p = .375$, $d_z = -0.10$. The three-way interaction between Target, Prime, and Condition was not significant, $F(1, 72) = 1.10$, $p = .297$, $\eta^2_p = .02$.

Discussion

In the current study we investigated whether a standard version and an adjusted version of the AAT procedure could be effective to change the evaluation of fear conditioned stimuli (neutral female faces paired with a loud unpleasant noise). Following fear conditioning, participants were asked to approach the CS+ and avoid the CS- using a joystick while either also approaching and avoiding neutral stimuli (pictures of a lamp and an umbrella) or while also approaching a positive picture (puppies) and avoiding a negative picture (a dead cat).

Our results can be summarized with two important findings. First, it appears that AAT can be effectively used to change conditioned fear and valence ratings. Effects were large (Cohen's $d_z = -0.98$ and 0.83 for fear and valence ratings, respectively). Of course, it should be

[3] As reported in our preregistration, we also analyzed the RTs of the AP task using a mixed-model approach. This analysis also did not yield the expected interaction between Condition (neutral AAT, valenced AAT) and Trial type (CS+ and positive target, CS+ and negative target, CS- and positive target, CS- and negative target), $F(3, 288) = 0.23$, $p = .878$. See the OSF page of this study for details regarding this analysis (https://osf.io/wphq8/).

noted that part of this effect can be due to updated expectations because there was no longer a negative sound during the AAT phase (i.e., extinction). Still, however, effects were large and got rid of almost all the negativity of the conditioned stimulus (see Figure 1). This is noteworthy, because many studies have previously found that extinction procedures are not very effective to change conditioned evaluations (Dirikx et al., 2004; Engelhard et al., 2014; Hofmann et al., 2010; Luck & Lipp, 2015). Hence, it appears that AAT may be a fairly effective way to counter conditioned negative conditioned valence and fear. Surprisingly, this conclusion runs counter the conclusions of Krypotos et al. (2015). This may be due to the different timing of collecting the evaluative ratings in our study and the study by Krypotos et al. (2015). Particularly, Krypotos and colleagues collected evaluative ratings at the end of the experiment, after a reinstatement phase (i.e., after 3 unannounced shock administrations). In contrast, we collected evaluative ratings of the CSs repeatedly throughout the experiment, including immediately after the AAT phase. Thereby, we presumably were able to capture participants' updated evaluation following the AAT procedure, whereas participants' in the study by Krypotos and colleagues probably relied on the whole experimental procedure or the reinstatement procedure to provide their evaluations. Indeed, in our own experiment we found that conditioned negative valence and fear ratings were restored to post-acquisition levels following the re-acquisition phase. Of note is also that we did not include physiological and behavioral measures of conditioned fear in our experiment, whereas Krypotos et al. (2015) did. Hence, we cannot confirm or disconfirm their findings on these measures (on which limited evidence for an effect of AAT was found).

A second important finding is that we did not find a stronger effect of the valenced AAT procedure, which goes against our prediction of strengthening the AAT effect due to approaching and avoiding strongly positive and negative stimuli. There are several possible reasons why our

adjusted AAT procedure did not more effectively change evaluations than the control version. First, a lack of statistical power could potentially be a problem. Descriptively, our valenced AAT procedure was slightly better to change conditioned valence and fear ratings after the AAT phase (see Figures 1 and 2). The sample we tested provides good statistical power (> .99) to detect large- (Cohen's $f = 0.40$) and medium-sized (Cohen's $f = 0.25$) interactions with condition. However, to detect smaller interaction effects (Cohen's $f \leq 0.10$) the statistical power of our sample was limited (< .43). Second, it might be that fear conditioning installs such strong shifts in valence that AAT may not be sufficient to change this conditioned valence. However, this interpretation seems unlikely because we observed clear changes in valence and fear from before to after AAT (see the previous paragraph). Third, perhaps the participants did not find the positive and negative picture sufficiently positive and negative (in contrast to the conditioned stimulus in Mertens et al., 2018), or they became habituated to these pictures throughout the AAT procedure. In future studies, this feature of our study could be improved by selecting a larger set of positive and negative pictures, or by having participants select positive and negative pictures themselves. Fourth, perhaps positive and negative pictures have only limited impact in the context of AAT, because the approach and avoidance movement already have strong evaluative components and non-focal stimuli cannot further strengthen the evaluative properties of the actions (note that in the study by Mertens et al., 2018, the effects of AAT on evaluations were reversed by the inclusion of non-focal stimuli, not strengthened).

For practical purposes (i.e., changing evaluations in clinical or marketing contexts) it seems that our adjusted version of the AAT procedure could be used, though its added value to the standard AAT procedure may be limited. Nonetheless, it would be interesting to see how our adapted procedure performs for changing the evaluations of neutral stimuli. It may also be the

case that in the context of a laboratory experiments in which a strongly aversive US (95 dB noise) was used, that the relative importance of executing approach-avoidance actions to determine evaluative responses is reduced. Hence, the dynamics of changing evaluations could be different in more neutral marketing contexts or in a clinical context where imminent threat is not present.

Finally, two limitations of our study should be noted. First, we did not include a sham-training or no-training control condition in our study. This complicates attributions of reduced valence and fear ratings to the AAT specifically. However, as mentioned above, note that prior research has found little evidence for reductions in conditioned valence without interventions or with standard extinction interventions. Second, implicit evaluations were only obtained at one time point (i.e., after the AAT intervention). Therefore, changes in implicit evaluations due to our intervention could not be captured (i.e., only relative differences between the two conditions in implicit evaluations could be assessed). Taking these considerations in mind, we conclude that a more standard and an adjusted version of the AAT procedure produced large shifts in conditioned negative evaluations and fear ratings. We did not find strong evidence for an advantage of our adjusted version of the AAT.

References

Becker, D., Jostmann, N. B., Wiers, R. W., & Holland, R. W. (2015). Approach avoidance training in the eating domain: Testing the effectiveness across three single session studies. *Appetite*, *85*(April 2016), 58–65. https://doi.org/10.1016/j.appet.2014.11.017

De Houwer, J. (2018). Propositional Models of Evaluative Conditioning. *Social Psychological Bulletin*, *13*(3). https://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853–869. https://doi.org/10.1037/0033-2909.127.6.853

Dirikx, T., Hermans, D., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2004). Reinstatement of extinguished conditioned responses and negative stimulus valence as a pathway to return of fear in humans. *Learning & Memory*, *11*(5), 549–554. https://doi.org/10.1101/lm.78004

Engelhard, I. M., Leer, A., Lange, E., & Olatunji, B. O. (2014). Shaking that icky feeling: Effects of extinction and counterconditioning on disgust-related evaluative learning. *Behavior Therapy*, *45*(5), 708–719. https://doi.org/10.1016/j.beth.2014.04.003

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological*

Bulletin, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gray, M. J., Litz, B. T., Hsu, J. L., & Lombardo, T. W. (2004). Psychometric Properties of the Life Events Checklist. *Assessment*, *11*(4), 330–341. https://doi.org/10.1177/1073191104269954

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Huijding, J., Field, A. P., De Houwer, J., Vandenbosch, K., Rinck, M., & van Oeveren, M. (2009). A behavioral route to dysfunctional representations: The effects of training approach or avoidance tendencies towards novel animals in children. *Behaviour Research and Therapy*, *47*(6), 471–477. https://doi.org/10.1016/j.brat.2009.02.011

Jones, C. R., Vilensky, M. R., Vasey, M. W., & Fazio, R. H. (2013). Approach behavior can mitigate predominately univalent negative attitudes: evidence regarding insects and spiders. *Emotion (Washington, D.C.)*, *13*(5), 989–996. https://doi.org/10.1037/a0033164

Kang, S., Vervliet, B., Engelhard, I. M., van Dis, E. A. M., & Hagenaars, M. A. (2018). Reduced return of threat expectancy after counterconditioning versus extinction. *Behaviour Research and Therapy*, *108*(June), 78–84. https://doi.org/10.1016/j.brat.2018.06.009

Krishna, A., & Eder, A. B. (2019). The influence of pre-training evaluative responses on approach-avoidance training outcomes. *Cognition and Emotion*, *0*(0), 1–14. https://doi.org/10.1080/02699931.2019.1568230

Krypotos, A.-M., Arnaudova, I., Effting, M., Kindt, M., & Beckers, T. (2015). Effects of Approach-Avoidance Training on the Extinction and Return of Fear Responses. *PLOS ONE*, *10*(7), e0131581. https://doi.org/10.1371/journal.pone.0131581

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Gainesville, Fl.

Luck, C. C., & Lipp, O. V. (2015). A potential pathway to the relapse of fear? Conditioned negative stimulus evaluation (but not physiological responding) resists instructed extinction. *Behaviour Research and Therapy*, *66*, 18–31. https://doi.org/10.1016/j.brat.2015.01.001

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Mertens, G., Van Dessel, P., & De Houwer, J. (2018). The contextual malleability of approach-avoidance training effects: approaching or avoiding fear conditioned stimuli modulates effects of approach-avoidance training. *Cognition and Emotion*, *32*(2), 341–349. https://doi.org/10.1080/02699931.2017.1308315

Mertens, G., Wagensveld, P., & Engelhard, I. M. (2019). Cue conditioning using a virtual spider discriminates between high and low spider fearful individuals. *Computers in Human Behavior*, *91*(October 2018), 192–200. https://doi.org/10.1016/j.chb.2018.10.006

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In *Communication and Persuasion* (Vol. 19, pp. 1–24). New York: Springer. https://doi.org/10.1007/978-1-4612-4964-1_1

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). Mind the gap: Increasing associations between the self and blacks with approach behaviors. *Journal of Personality and Social Psychology*, *100*(2), 197–210. https://doi.org/10.1037/a0022159

Van Dessel, P., Hughes, S., & De Houwer, J. (2018). How Do Actions Influence Attitudes? An Inferential Account of the Impact of Action Performance on Stimulus Evaluation.

Personality and Social Psychology Review, 108886831879573. https://doi.org/10.1177/1088868318795730

van Uijen, S., van den Hout, M., & Engelhard, I. (2015). Active Approach Does not Add to the Effects of in Vivo Exposure. *Journal of Experimental Psychopathology*, *6*(1), 112–125. https://doi.org/10.5127/jep.042014

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining Automatic Action Tendencies Changes Alcoholic Patients' Approach Bias for Alcohol and Improves Treatment Outcome. *Psychological Science*, *22*(4), 490–497. https://doi.org/10.1177/0956797611400615

Zbozinek, T. D., Hermans, D., Prenoveau, J. M., Liao, B., & Craske, M. G. (2015). Post-extinction conditional stimulus valence predicts reinstatement fear: Relevance for long-term outcomes of exposure therapy. *Cognition and Emotion*, *29*(4), 654–667. https://doi.org/10.1080/02699931.2014.930421

Supplementary Materials: "Does approaching puppies and avoiding a dead cat improve the effectiveness of approach-avoidance training for changing the evaluation of feared stimuli?"

Table 1. Mean (SD) valence ratings throughout the experiment.

	Pre-acq	Post-acq	Post-AAT	Post-reacq
Valenced AAT				
(n = 36)				
CS+	43.83	29.69	46.39	30.32
	(14.42)	(13.94)	(14.31)	(14.95)
CS-	44.00	52.47	45.31	52.86
	(12.85)	(14.67)	(10.37)	(10.51)
Standard AAT				
(n = 38)				
CS+	42.16	30.74	43.84	29.26
	(10.11)	(14.69)	(11.14)	(16.81)
CS-	42.47	53.71	50.66	55.29
	(11.53)	(14.74)	(12.00)	(14.90)

Table 2. Mean (SD) fear ratings throughout the experiment.

	Pre-acq	Post-acq	Post-AAT	Post-reacq
Valenced AAT				
(n = 36)				
CS+	26.47	51.92	30.39	49.53
	(24.66)	(27.67)	(23.41)	(29.51)
CS-	24.61	16.11	25.83	17.64
	(23.77)	(21.20)	(25.13)	(18.93)
Standard AAT				
(n = 38)				
CS+	31.55	65.97	41.32	62.84
	(24.53)	(24.43)	(23.05)	(27.25)
CS-	34.24	19.37	23.58	20.16
	(24.72)	(18.51)	(21.10)	(19.31)

The research reported in this paper was funded by a NWO VICI grant (grant number: 453-15-005) and a Postdoctoral fellowship of the Scientific Research Foundation, Flanders (FWO-Vlaanderen). We declare no conflict of interest with regard to the preparation of this manuscript and confirm that our manuscript is an original contribution which is not under review or published anywhere else.